Toward negotiable reinforcement learning: shifting priorities in Pareto optimal sequential decision-making

Andrew Critch[*][†]

January 4, 2017

Abstract

Existing multi-objective reinforcement learning (MORL) algorithms do not account for objectives that arise from players with differing beliefs. Concretely, consider two players with different beliefs and utility functions who may cooperate to build a machine that takes actions on their behalf. A representation is needed for how much the machine's policy will prioritize each player's interests over time. Assuming the players have reached common knowledge of their situation, this paper derives a recursion that any Pareto optimal policy must satisfy. Two qualitative observations can be made from the recursion: the machine must (1) use each player's own beliefs in evaluating how well an action will serve that player's utility function, and (2) shift the relative priority it assigns to each player's expected utilities over time, by a factor proportional to how well that player's beliefs predict the machine's inputs. Observation (2) represents a substantial divergence from naïve linear utility aggregation (as in Harsanyi's utilitarian theorem, and existing MORL algorithms), which is shown here to be inadequate for Pareto optimal sequential decision-making on behalf of players with different beliefs.

1 Introduction

It has been argued that the first AI systems with generally super-human cognitive abilities will play a pivotal decision-making role in directing the future of civilization (Bostrom, 2014). If that is the case, an important question will arise: *Whose values will the first super-human AI systems serve?* Since safety is a crucial consideration in developing such systems, assuming the institutions building them come to understand the risks and the time investments needed to address them (Baum, 2016), they will have a large incentive to cooperate in

[*] Machine Intelligence Research Institute

[†] UC Berkeley, Center for Human Compatible AI

their design rather than racing under time-pressure to build competing systems (Armstrong, Bostrom, and Shulman, 2016).

Therefore, consider two nations—allies or adversaries—who must decide whether to cooperate in the deployment of an extremely powerful AI system. Implicitly or explicitly, the resulting system would have to strike compromises when conflicts arise between the wishes of those nations. How can they specify the degree to which that system would be governed by the distinctly held principles of each nation? More mundanely, suppose a couple purchases a domestic robot. How should the robot strike compromises when conflicts arise between the commands of its owners?

It is already an interesting and difficult problem to robustly align an AI system's values with those of a single *single* human (or a group of humans in close agreement). Inverse reinforcement learning (IRL) (S. Russell, 1998) (Ng and S. J. Russell, 2000) (Abbeel and Ng, 2004) and cooperative inverse reinforcement learning (CIRL) (Hadfield-Menell et al., 2016) represent successively realistic early approaches to this problem. But supposing some adequate solution eventually exists for aligning the values of a machine intelligence with a single human decision-making unit, how should the values of a system serving *multiple* decision-makers be "aligned"?

One might hope to specify some extremely compelling ethical principle that everyone would immediately accept. Realistically, however, disagreements will always exist. Consider the general case of two parties—perhaps states, companies, or individuals—who might cooperatively build or purchase an AI system to serve them both.[1] If the parties cannot reach sufficient agreement as to what policy the AI should follow, cooperation may be less attractive than obtaining separate AI systems, one for each party. At the individual level, non-cooperation could mean domestic disputes between domestic robots. At the state level, it could mean an arms race between nations competing under time pressure to develop ever more powerful militarized AI systems, affording each nation less time to ensure the safety and validity of their respective systems.

Unless the prospect of cooperative AI ownership is made sufficiently attractive to the separate parties, the question of whose values the cooperatively owned system "ought" to serve is moot: the parties will fall back on non-cooperative strategies—perhaps obtaining separate machines that will compete with each other—and the jointly owned system will not exist in the first place. In addition, if the process of bargaining over the policy of a cooperatively owned system is difficult or complicated, the players are more likely to end negotiations and default to non-cooperative strategies.

Conversely, if bargaining is made easier, players are more likely to reach cooperation. The purpose of this paper is to begin formalizing the problem of negotiating over the policy of a machine intelligence, and to exhibit some early findings as to the nature of *Pareto optimal policies*—policies which cannot be improved for one player without sacrifice by another—with the eventual aim

[1] The results of this paper all generalize directly from 2 to n players, but for concreteness of exposition, the two-player case is prioritized.

of making cooperative outcomes easier to formulate, more attractive, and more likely to obtain.

Outline. The paper is organized as follows. Section 2 briefly outlines some standard choices of notation. Section 3 formalizes the problem of obtaining a Pareto optimal policy for two distinct parties with common knowledge of distinct priors, derives a recursion that any such policy must follow, and contrasts that recursion with a more naïve "just add up a linear combination of the utility functions" approach. The recursion implies two main qualitative insights about how a Pareto optimal policy, pursuant to a common-knowledge difference in opinion, must behave over time: (1) such a policy must (explicitly or implicitly) use each player's own beliefs in evaluating how well an action will serve that player's utility function, and (2) it must shift the relative priority it places on each player's expected utilities over time, by a factor proportional to how well that player's beliefs predict the machine's inputs. Section 4 provides some further interpretation of these implications, in terms of bet-settling and moral realism. Section 5 outlines subsequent work expected to be useful for enabling cooperative AI deployment in the future. Finally, Section 6 provides concluding remarks targeted at readers who have finished the full paper.

1.1 Related work

Social choice theory. The whole of social choice theory and voting theory may be viewed as an attempt to specify an agreeable formal policy to enact on behalf of a group. Harsanyi's utility aggregation theorem (Harsanyi, 1980) suggests one form of solution: maximizing a linear combination of group members' utility functions. The present work shows that this solution is inappropriate when players have different beliefs, and Theorem 8 may be viewed as an extension of Harsanyi's form that accounts simultaneously for differing priors and the prospect of future observations. Indeed, Harsanyi's form follows as a direct corollary of Theorem 8 when players do share the same beliefs (Corollary 9).

Bargaining theory. The formal theory of bargaining, as pioneered by Nash (Nash, 1950) and carried on by authors such as Myerson (Myerson, 1979) (Myerson, 2013) and Satterthwaite (Myerson and Satterthwaite, 1983), is also extremely topical. Future investigation in this area might be aimed at generalizing their work to sequential decision-making settings, and this author recommends a focus on research specifically targeted at resolving conflicts.

Multi-agent systems. There is ample literature examining multi-agent systems using sequential decision-making models. Shoham and Leyton-Brown (2008) survey various models of multiplayer games using an MDP to model each agent's objectives. Chapter 9 of the same text surveys social choice theory, but does not account for sequential decision-making.

Zhang and Shah (2014) may be considered a sequential decision-making approach to social choice: they use MDPs to represent the decisions of players in a competitive game, and exhibit an algorithm for the players that, if followed, arrives at a Pareto optimal Nash equilibrium satisfying a certain fairness criterion. Among the literature surveyed here, that paper is the closest to the present work in terms of its intended application: roughly speaking, achieving mutually desirable outcomes via sequential decision-making. However, that work is concerned with an ongoing interaction between the players, rather than selecting a policy for a single agent to follow as in this paper.

Multi-objective sequential decision-making. There is also a good deal of work on Multi-Objective Optimization (MOO) (Tzeng and Huang, 2011), including for sequential decision-making, where solution methods have been called Multi-Objective Reinforcement Learning (MORL). For instance, Gbor, Kalmr, and Szepesvri (1998) introduce a MORL method called Pareto Q-learning for learning a set of a Pareto optimal polices for a Multi-Objective MDP (MOMDP). Soh and Demiris (2011) define Multi-Reward Partially Observable Markov Decision Processes (MR-POMDPs), and use use genetic algorithms to produce non-dominated sets of policies for them. Roijers, Whiteson, and Oliehoek (2015) refer to the same problems as Multi-objective POMDPS (MOPOMDPs), and provide a bounded approximation method for the optimal solution set for all possible weightings of the objectives. Wang (2014) surveys MORL methods, and contributes Multi-Objective Monte-Carlo Tree Search (MOMCTS) for discovering multiple Pareto optimal solutions to a multi-objective optimization problem. Wray and Zilberstein (2015) introduce Lexicographic Partially Observable Markov Decision Process (LPOMDPs), along with two accompanying solution methods.

However, none of these or related works address scenarios where the objectives are derived from players with differing beliefs, from which the priority-shifting phenomenon of Theorem 8 arises. Differing beliefs are likely to play a key role in negotiations, so for that purpose, the formulation of multi-objective decision-making adopted here is preferable.

2 Notation

The reader is invited to skip this section and refer back as needed; an effort has been made to use notation that is intuitive and fairly standard, following Pearl (2009), Hutter (2003), and Orseau and Ring (2012).

Random variables are denoted by uppercase letters, e.g., S_1, and lowercase letters, e.g., s_1, are used as indices ranging over the values of a variable, as in the equation

$$\mathbb{E}[S_1] = \sum_{s_1} \mathbb{P}(s_1) \cdot s_1.$$

Sequences are denoted by overbars, e.g., given a sequence $(s_1, \ldots, s_n)$, $\bar{s}$

stands for the whole sequence. Subsequences are denoted by subscripted inequalities, so e.g., $s_{<4}$ stands for (s_1, s_2, s_3), and $s_{\leq 4}$ stands for (s_1, s_2, s_3, s_4).

3 Two agents building a third

Consider, informally, a scenario wherein two players — perhaps individuals, companies, or states — are considering cooperating to build or otherwise obtain a machine that will then interact with an environment on their behalf.[2] In such a scenario, the players will tend to bargain for "how much" the machine will prioritize their separate interests, so to begin, we need some way to quantify "how much" each player is prioritized.

For instance, one might model the machine as maximizing the expected value, given its observations, of some utility function U of the environment that equals a weighted sum

$$w^1 U^1 + w^2 U^2 \tag{1}$$

of the players' individual utility functions U^1 and U^2, as Harsanyi's social aggregation theorem (Harsanyi, 1980) recommends. Then the bargaining process could focus on choosing the values of the weights w^i.

However, this turns out to be a bad idea. As we shall see in Proposition 10, this solution form is not generally compatible with Pareto optimality when agents have different beliefs. Harsanyi's setting does not account for agents having different priors, nor for decisions being made sequentially, after future observations. In such a setting, we need a new form of solution, exhibited here along with a recursion that characterizes optimal solutions by a process analogous to, but meaningfully different from, Bayesian updating.

3.1 A POMDP formulation

Let us formalize the machine's decision-making situation using the structure of a Partially Observable Markov Decision Process (POMDP), as depicted by the Bayesian network in Figure 1. (See S. Russell et al. (2003) for an introduction to POMDPs, and Darwiche (2009) for an introduction to Bayesian networks.)

At each point in time i, the machine will have a policy π_i that for each possible sequence of observations $o_{\leq i}$ and past actions $a_{<i}$, returns a distribution $\pi_i(- \mid o_{\leq i} a_{<i})$ on actions a_i, which will then be used to generate an action a_i with probability $\pi(a_i \mid o_{\leq i} a_{<i})$. In Figure 1, the part of the Bayes net governed by the machine's policy is highlighted in green.

Common knowledge assumptions. It is assumed that the players will have common knowledge of the policy $\pi = (\pi_1, \ldots, \pi_n)$ they select for the machine to implement, but that the players may have different beliefs about how the

[2] The results here all generalize from two players to n players being combined successively in any order, but for clarity of exposition, the two person case is prioritized.

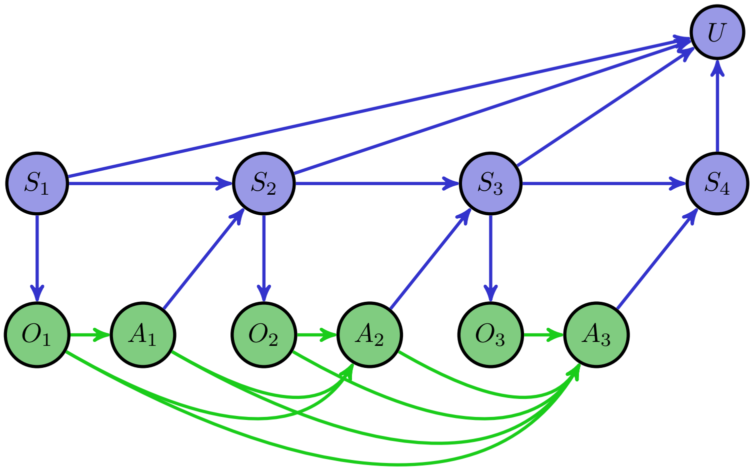

Figure 1: A POMDP of length $n = 3$

environment works, and of course different utility functions. It is also assumed that the players have common knowledge of one another's posterior.

This assumption is critical. During a bargaining process, one should expect players' beliefs to update in response to one another's behavior. Assuming common knowledge of posteriors means that the players have reached an equilibrium where, each knowing what the other believes, does not wish to further update her own beliefs.[3]

We encode each player j's outlook as a POMDP, $D^j = (\mathcal{S}^j, \mathcal{A}, T^j, U^j, \mathcal{O}, \Omega^j, n)$, which simultaneously represents that player's beliefs about the environment, and the player's utility function.

- $\mathcal{S}^j$ represents a set of possible states s of the environment,
- $\mathcal{A}$ represents the set of possible actions a available to the machine,
- T^j represents the conditional probabilities player j believes will govern the environment state transitions, i.e., $\mathbb{P}^j(s_{i+1} \mid s_i a_i)$,
- U^j represents player j's utility function from sequences of environmental states $(s_1, \ldots, s_n)$ to $\mathbb{R}$; for the sake of generality, U^j is *not assumed* to be additive over time, as reward functions often are,
- $\mathcal{O}$ represents the set of possible observations o of the machine,

[3] Future work should design solutions for facilitating the process of attaining common knowledge, or to obviate the need to assume it.

- Ω^j represents the conditional probabilities player j believes will govern the machine's observations, i.e., $\mathbb{P}^j(o_i \mid s_i)$, and
- n is the number of time steps.

Thus, player j's subjective probability of an outcome $(\bar{s}, \bar{o}, \bar{a})$, for any $\bar{s} \in (\mathcal{S}^j)^n$, is given by a probability distribution $\mathbb{P}^j$ that takes π as a parameter:

$$\mathbb{P}^j(\bar{s}, \bar{o}, \bar{a}; \pi) := \mathbb{P}^j(s_1) \cdot \prod_{i=1}^{n} \mathbb{P}^j(o_i \mid s_i)\, \pi(a_i \mid o_{\leq i} a_{<i})\, \mathbb{P}^j(s_{i+1} \mid s_i a_i) \qquad (2)$$

As such, the POMDPs D^1 and D^2 are "compatible" in the following sense:

Definition 1 (Compatible POMDPs)**.** *We say that two POMDPs, D^1 and D^2, are* compatible *if any policy for one may be viewed as a policy for the other, i.e., they have the same set of actions $\mathcal{A}$ and observations $\mathcal{O}$, and the same number of time steps n.*

3.2 Pareto optimal policies

In this context, where a policy π may be evaluated relative to more than one POMDP, we use superscripts to represent which POMDP is governing the probabilities and expectations, e.g.,

$$\mathbb{E}^j[U^j; \pi] := \sum_{\bar{s} \in (\mathcal{S}^j)^n} \mathbb{P}^j(\bar{s}; \pi) U^j(\bar{s})$$

represents the expectation in D^j of the utility function U^j, assuming policy π is followed.

Definition 2 (Pareto optimal policies)**.** *A policy π is* Pareto optimal *for a compatible pair of POMDPs (D^1, D^2) if for any other policy π', either*

$$\mathbb{E}^1[U^1; \pi] \geq \mathbb{E}^1[U^1; \pi'] \quad \textit{or} \quad \mathbb{E}^2[U^2; \pi] \geq \mathbb{E}^2[U^2; \pi'].$$

It is assumed that, during negotiation, the players will be seeking a Pareto optimal policy for the machine to follow, relative to the POMDPs D^1 and D^2 describing each player's outlook.

Policy mixing assumption. It is also assumed that during the agent's first action (or before it), the agent has the ability to generate and store some random numbers in the interval $[0, 1]$, called a random seed, that will not affect the environment except through other features of its actions. Then, given any two policies π and π' and a scalar $p \in [0, 1]$ we may construct a third policy,

$$p\pi + (1 - p)\pi',$$

that decides with probability p (before receiving any inputs) to use policy π for generating all of its future actions, and otherwise uses policy π'. (This is a "once and for all" decision; the agent does not flip-flop between π and π' once the decision is made.) Mixtures of more than two policies are defined similarly. With this formalism, whenever $\sum_k \alpha_k = 1$ and each $\alpha_k \geq 0$, we have

$$\mathbb{E}^j \left[U^j; \sum_k \alpha_k \pi_k \right] = \sum_k \alpha_k \mathbb{E}^j[U^j; \pi_k]. \tag{3}$$

Lemma 3. *A policy π is Pareto optimal to players 1 and 2 if and only if there exist weights $w^1, w^2 \geq 0$ with $w_1 + w_2 = 1$ such that*

$$\pi \in \underset{\pi^* \in \Pi}{\operatorname{argmax}} \left(w^1 \mathbb{E}^1[U^1; \pi^*] + w^2 \mathbb{E}^2[U^2; \pi^*] \right) \tag{4}$$

Proof. The mixing assumption gives the space of policies Π the structure of a convex space that the maps $\mathbb{E}^j[U^j; -]$ respect by Equation 3. This ensures that the image of the map $f : \Pi \to \mathbb{R}^2$ given by

$$f(\pi) := \left(\mathbb{E}^1[U^1; \pi],\ \mathbb{E}^2[U^2; \pi] \right)$$

is a closed, convex polytope. As such, a point (x, y) lies on the Pareto boundary of image(f) if and only if there exist nonnegative weights (w^1, w^2), not both zero, such that

$$(x, y) \in \underset{(x^*, y^*) \in \text{image}(f)}{\operatorname{argmax}} \left(w^1 x^* + w^2 y^* \right)$$

After normalizing $w^1 + w^2$ to equal 1, this implies the result. □

3.3 A reprioritization mechanism that resembles Bayesian updating

We shall soon see that any Pareto optimal policy π must favor, as time progresses, optimizing the *utility* of whichever player's *beliefs* were a better predictor of the machine's inputs. This phenomenon turns out to algebraically resemble Bayesian updating, but is quite different in its meaning. Nonetheless, it is most easily shown to occur by a precise analogy to Bayesian updating in a third POMDP constructed from the outlooks of players 1 and 2, as follows.

For any weights, $w^1, w^2 \geq 0$ with $w^1 + w^2 = 1$, we define a new POMDP that works by flipping a (w^1, w^2)-weighted coin, and then running D^1 or D^2 thereafter, according to the coin flip. Explicitly,

Definition 4 (POMDP mixtures). *Let D^1 and D^2 be compatible POMDPs, with parameters $D^j = (\mathcal{S}^j, \mathcal{A}, T^j, U^j, \mathcal{O}, \Omega^j, n)$. Define a new POMDP compatible both, denoted $D = w^1 D^1 + w^2 D^2$, with parameters $D^j = (\mathcal{S}, \mathcal{A}, T, U, \mathcal{O}, \Omega, n)$, as follows:*

- $\mathcal{S} := \{(j, s) \mid j \in \{1, 2\}, s \in \mathcal{S}^j\}$,

- *The environmental transition probabilities T are given by*

$$\mathbb{P}\left((j, s_1)\right) := w^j \cdot \mathbb{P}^j(s_1)$$

for any initial state $s_1 \in \mathcal{S}^j$, and thereafter,

$$\mathbb{P}\Big((j', s_{i+1}) \mid (j, s_i), a_i\Big) := \begin{cases} \mathbb{P}^j\left(s_{i+1} \mid s_i a_i\right) & \textit{if } j' = j \\ 0 & \textit{if } j' \neq j \end{cases}$$

Hence, the value of j will be constant over time, so a full history for the environment may be represented by a pair

$$(j, \bar{s}) \in \{1\} \times (\mathcal{S}^1)^n \cup \{2\} \times (\mathcal{S}^2)^n.$$

Let B denote the boolean random variable that equals whichever constant value of j obtains, so then

$$\mathbb{P}(B = j) = w^j$$

- *The utility function U is given by*

$$U(j, \bar{s}) := U^j(\bar{s})$$

- *The observation probabilities Ω are given by*

$$\mathbb{P}\left(o_i \mid (j, s_i)\right) := \mathbb{P}(B = j) \cdot \mathbb{P}^j(o_i \mid s_i)$$

In particular, the policy does not observe directly whether $j = 1$ or $j = 2$.

The POMDP mixture $D = w^1 D^1 + w^2 D^2$ can be depicted with a Bayes net by adding an additional environmental node for B in the diagram of D^1 and D^2 (see Figure 2). Indeed, given any policy π, the expected payoff of π in $w^1 D^1 + w^2 D^2$ is exactly

$$\begin{aligned} &\mathbb{P}(B = 1) \cdot \mathbb{E}(U \mid B = 1; \pi) + \mathbb{P}(B = 2) \cdot \mathbb{E}(U \mid B = 2; \pi) \\ &= w^1 \mathbb{E}^2(U^1; \pi) + w^2 \mathbb{E}^2(U^2; \pi) \end{aligned}$$

Therefore, using the above definitions, Lemma 3 may be restated in the following equivalent form:

Lemma 5. *Given a pair (D^1, D^2) of compatible POMDPs, a policy π is Pareto optimal for that pair if and only if there exist weights w^j such that π is an optimal policy for the single POMDP given by $w^1 D^1 + w^2 D^2$.*

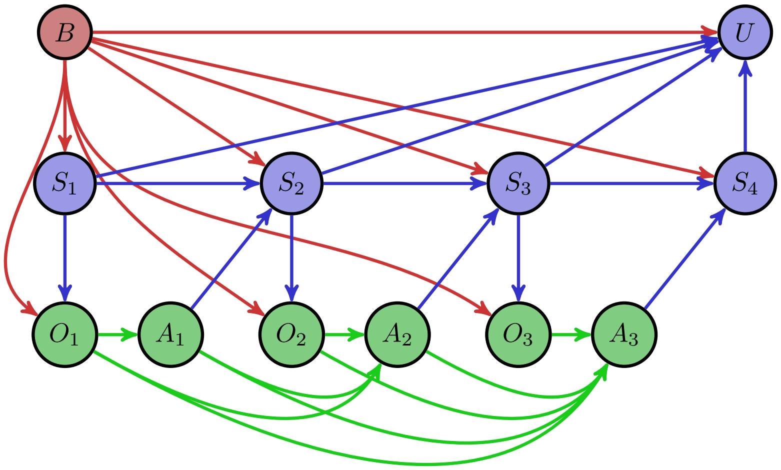

Figure 2: A POMDP (mixture) of length $n = 3$ initialized by a Boolean B

3.4 A recursive Pareto optimality condition

Expressed in the form of Equation 4, it might not be clear how a Pareto optimal policy makes use of its observations over time, aside from storing them in memory. For example, is there any sense in which the machine carries "beliefs" about the environment that it "updates" at each time step? Lemma 5 allows us to answer this and related questions by translating theorems about single POMDPs into theorems about compatible pairs of POMDPs.

If π is an optimal policy for a single POMDP, at any time step i, optimality of the action distribution $\pi_i(- \mid o_{\leq i}a_{<i})$ can be characterized without reference to the previous policy components $(\pi_1, \ldots, \pi_{i-1})$, nor to $\pi_i(- \mid o'_{\leq i}a'_{<i})$ for any alternate history $o'_{\leq i}a'_{<i}$.[4] To express this claim in an equation, Pearl's "$do()$" notation (Pearl, 2009) comes in handy:

Definition 6 ("do" notation)**.**

$$\mathbb{P}^j(\bar{o} \mid do(\bar{a})) := \sum_{\bar{s} \in (\mathcal{S}^j)^n} \mathbb{P}^j(s_1) \cdot \prod_{i=1}^{n} \mathbb{P}^j(o_i \mid s_i)\, \mathbb{P}^j(s_{i+1} \mid s_i a_i)$$

This expression is the same as the probability of $(\bar{o}, \bar{a})$ when π is the constant policy that places probability 1 on the action sequence $\bar{a}$.

Proposition 7 (Classical separability)**.** *If D is a POMDP described by conditional probabilities $\mathbb{P}(- \mid -)$ and utility function U (as in Equation 2), then a*

[4] This fact can used to justify why the "sunk cost" fallacy is indeed a fallacy.

policy π is optimal for D if and only if for each time step i and each observation/action history $o_{\leq i}a_{<i}$, the action distribution $\pi_i(- \mid o_{\leq n}a_{<n})$ satisfies the following backward recursion:

$$\pi_i(- \mid o_{\leq i}a_{<i}) \in \underset{\alpha \in \Delta A}{\operatorname{argmax}} \Big(\\ \mathbb{P}(o_{\leq i} \mid do(a_{<i})) \cdot \mathbb{E}[U \mid o_{\leq i}a_{<i};\ a_n \sim \alpha;\ \pi_{i+1}, \ldots, \pi_n] \Big)$$

This characterization of $\pi_i(o_{\leq i}a_{<i})$ does not refer to $\pi_1, \ldots, \pi_{i-1}$, nor to $\pi_i(o'_{\leq i}a'_{<i})$ for any alternate history $o'_{\leq i}a'_{<i}$.

Proof. This is a standard property of POMDP solutions. □

It turns out that Pareto optimality can be characterized in a similar way by backward recursion from the final time step. The resulting recursion reveals a pattern in how the weights on the players' conditionally expected utilities must change over time, which is the main result of this paper:

Theorem 8 (Pareto optimal policy recursion)**.** *Given a pair (D^1, D^2) of compatible POMDPs of length n, a policy π is Pareto optimal if and only if its components π_i for $i \leq n$ satisfy the following backward recursion for some pair of weights $w^1, w^2 \geq 0$ with $w^1 + w^2 = 1$:*

$$\pi^i(- \mid o_{\leq i}a_{<i}) \in \underset{\alpha \in \Delta A}{\operatorname{argmax}} \Big(\\ w^1\mathbb{P}^1\left(o_{\leq i} \mid do(a_{<i}); \pi_{i+1}, \ldots, \pi_n\right) \cdot \mathbb{E}^1[U^1 \mid o_{\leq i}a_{<i}a_i;\ a_i \sim \alpha;\ \pi_{i+1}, \ldots, \pi_n] \\ + w^2\mathbb{P}^2\left(o_{\leq i} \mid do(a_{<i}); \pi_{i+1}, \ldots, \pi_n\right) \cdot \mathbb{E}^2[U^2 \mid o_{\leq i}a_{<i}a_i;\ a_i \sim \alpha;\ \pi_{i+1}, \ldots, \pi_n] \Big)$$

In words, to achieve Pareto optimality, the machine must

1. *use each player's own beliefs when estimating the degree to which a decision favors that player's utility function, and*
2. *shift the relative priorities of the players' expected utilities in the machine's decision objective over time, by a factor proportional to how well the players predict the machine's inputs.*

Proof. By Lemma 5, the Pareto optimality of π for (D^1, D^2) is equivalent to its classical optimality for $w^1D^1 + w^2D^2$ for some (w^1, w^2), which by Proposition 7 is equivalent to satisfying the following backward recursion (writing $\mathbb{P}$ and $\mathbb{E}$ for probabilities and expectations in $w^1D^1 + w^2D^2$):

$$\pi^i(- \mid o_{\leq i}a_{<i}) \in \underset{\alpha \in \Delta A}{\operatorname{argmax}} \Big(\\ \mathbb{P}(B = 1) \cdot \mathbb{P}\left(o_{\leq i} \mid do(a_{<i}); \pi_{i+1}, \ldots, \pi_n\right) \cdot \mathbb{E}[U \mid o_{\leq i}a_{<i}a_i;\ a_i \sim \alpha;\ \pi_{i+1}, \ldots, \pi_n] \\ + \mathbb{P}(B = 2) \cdot \mathbb{P}\left(o_{\leq i} \mid do(a_{<i}); \pi_{i+1}, \ldots, \pi_n\right) \cdot \mathbb{E}[U \mid o_{\leq i}a_{<i}a_i;\ a_i \sim \alpha;\ \pi_{i+1}, \ldots, \pi_n] \Big).$$

By Definition 4, the expression inside the argmax equals

$$w^1\mathbb{P}^1\left(o_{\le i} \mid do(a_{<i}); \pi_{i+1}, \ldots, \pi_n\right) \cdot \mathbb{E}^1[U^1 \mid o_{\le i}a_{<i}a_i;\ a_i \sim \alpha;\ \pi_{i+1}, \ldots, \pi_n]$$
$$+w^2\mathbb{P}^2\left(o_{\le i} \mid do(a_{<i}); \pi_{i+1}, \ldots, \pi_n\right) \cdot \mathbb{E}^2[U^2 \mid o_{\le i}a_{<i}a_i;\ a_i \sim \alpha;\ \pi_{i+1}, \ldots, \pi_n]$$

hence the result. □

When the players have the same beliefs, they aways assign the same probability to the machine's inputs, so the weights on their respective expectations do not change over time. In this case, Harsanyi's utility aggregation formula is recovered as a special instance:

Corollary 9 (Harsanyi's utility aggregation formula)**.** *Suppose that players 1 and 2 share the same beliefs about the environment, i.e., the pair (D^1, D^2) of compatible POMDPs agree on all parameters except the players' utility functions $U^1 \neq U^2$. Then a policy π is Pareto optimal if and only if there exist weights $w^1, w^2 \geq 0$ with $w^1 + w^2 = 1$ such that for $i \leq n$, π_i satisfies*

$$\pi^i(- \mid o_{\le i}a_{<i}) \in \operatorname*{argmax}_{\alpha \in \Delta A}\left(\mathbb{E}[w^1U^1 + w^2U^2] \mid o_{\le i}a_{<i}a_i;\ a_i \sim \alpha;\ \pi_{i+1}, \ldots, \pi_n]\right)$$

where $\mathbb{E} = \mathbb{E}^1 = \mathbb{E}^2$ denotes the shared expectations of both players.

Proof. Setting $\mathbb{E} = \mathbb{E}^1 = \mathbb{E}^2$ in Theorem 8, factoring out the common coefficient $\mathbb{P}^1\left(o_{\le i} \mid do(a_{<i}); \pi_{i+1}, \ldots, \pi_n\right) = \mathbb{P}^2\left(o_{\le i} \mid do(a_{<i}); \pi_{i+1}, \ldots, \pi_n\right)$, and applying linearity of expectation yields the result. □

3.5 Comparison to naïve utility aggregation

To see the necessity of the $\mathbb{P}^j$ terms that shift the expectation weights in Theorem 8 over time, let us compare it with the behavior of an alternative optimization criterion that maximizes a fixed linear combination of expectations.

A cake-splitting scenario. The parameters of this scenario are laid out in Table 1, and described as follows:

Alice (Player 1) and Bob (Player 2) are about to be presented with a cake which they can choose to split in half to share, or give entirely to one of them. They have (built or purchased) a robot that will make the cake-splitting decision on their behalf. Alice's utility function returns 0 if she gets no cake, 20 if she gets half a cake, or 30 if she gets a whole cake. Bob's utility function works similarly.

However, Alice and Bob have slightly different beliefs about how the environment works. They both agree on the state of the environment that the robot will encounter at first: a room with a cake in it (S_1 = "cake"). But Alice and Bob have different predictions about how the robot's sensors will perceive the cake: Alice thinks that when the robot perceives the cake, it is 90%

likely to appear with a red tint ($O_1 =$ "red"), and 10% likely to appear with a green tint ($O_1 =$ "green"), whereas Bob believes the exact opposite. In either case, upon seeing the cake, the robot will either give Alice the entire cake ($A_1 = S_1 =$ (all, none)), split the cake half-and-half ($A_1 = S_1 =$ (half, half)), or give Bob the entire cake ($A_1 = S_1 =$ (none, all)). Moreover, Alice and Bob have common knowledge of all these facts.

S_1	O_1	$\mathbb{P}^1(O_1 \mid S_1)$	$\mathbb{P}^2(O_1 \mid S_1)$	$A_1 = S_1$	U^1	U^2
cake	red	90%	10%	(all, none)	30	0
				(half, half)	20	20
				(none, all)	0	30
	green	10%	90%	(all, none)	30	0
				(half, half)	20	20
				(none, all)	0	30

Table 1: An example scenario wherein a Pareto optimal policy undergoes priority shifting

Now, consider the following Pareto optimal policy that favors Alice (Player 1) when O_1 is red, and Bob (Player 2) when O_1 is green:

$$\hat{\pi}(- \mid \text{red}) = 100\%(\text{all, none})$$
$$\hat{\pi}(- \mid \text{green}) = 100\%(\text{none, all})$$

This policy can be viewed intuitively as a bet between Alice and Bob about the value of O_1, and is highly appealing to both players:

$$\mathbb{E}^1[U^1; \hat{\pi}] = 90\%(30) + 10\%(0) = 27$$
$$\mathbb{E}^2[U^2; \hat{\pi}] = 10\%(0) + 90\%(30) = 27$$

In particular, $\hat{\pi}$ is more appealing to both Alice and Bob than an agreement to deterministically split the cake (half, half). However,

Proposition 10. *The Pareto optimal strategy $\hat{\pi}$ above cannot be implemented by any machine that naïvely maximizes a fixed-over-time linear combination of the conditionally expected utility of the two players, i.e., by any policy π satisfying*

$$\pi(- \mid o_1) \in \operatorname*{argmax}_{\alpha \in \Delta A} \Big(r \cdot \mathbb{E}^1[U^1 \mid o_1; a_1 \sim \alpha] + (1 - r) \cdot \mathbb{E}^2[U^2 \mid o_1; a_1 \sim \alpha] \Big) \quad (5)$$

for some fixed $r \in [0, 1]$. Moreover, every such policy π is strictly worse than $\hat{\pi}$ in expectation to one of the players.

This proposition is relatively unsurprising when one considers the policy $\hat{\pi}$ intuitively as a bet-settling mechanism, and that the nature of betting is to favor different preferences based on future observations. However, to be sure of this impossibility claim, one must rule out the possibility that the $\hat{\pi}$ could be implemented by having the machine choose which element of the argmax in Equation 5 to use based on whether the cake appears red or green.

Proof of Proposition 10. Suppose π is any policy satisfying Equation 5 for some fixed r, and consider the following cases for r:

1. If $r < 1/3$, then π must satisfy

$$\pi(- \mid o_1) = 100\%(\text{none, all}).$$

Here, $\mathbb{E}^1[U^1; \pi] = 0 < 27$, so π is strictly worse than $\hat{\pi}$ in expectation to Alice.

2. If $r = 1/3$, then π must satisfy

$$\pi(- \mid o_1) = q(o_1)(\text{none, all}) + (1 - q(o_1))(\text{half, half})$$

for some $q(o_1) \in [0, 1]$ depending on o_1. Here, $\mathbb{E}^1[U^1; \pi] \leq 20 < 27$ (with equality when $q(\text{red}) = q(\text{green}) = 1$), so π is strictly worse than $\hat{\pi}$ in expectation to Alice.

3. If $1/3 < r < 2/3$, then π must satisfy

$$\pi(- \mid o_1) = 100\%(\text{half, half})$$

Here, $\mathbb{E}^1[U^1; \pi] = \mathbb{E}^2[U^2; \pi] = 20 < 27$, so π is strictly worse than $\hat{\pi}$ in expectation to both Alice and Bob.

The remaining cases, $r = 2/3$ and $r > 2/3$, are symmetric to the first two, with Bob in place of Alice and (none, all) in place of (all, none). □

4 Interpretations

Theorem 8 shows that a Pareto optimal policy must tend, over time, toward prioritizing the expected *utility* of whichever player's *beliefs* best predict the machine's inputs better. From some perspectives, this is a little counterintuitive: not only must the machine gradually place more predictive weight on whichever player's prior is a better predictor, but it must reward that player by attending more to her utility function as well. This behavior is not an assumption, but rather is forced to occur by Pareto optimality. The players *must* agree to this pattern of shifting priority over time, or else they will leave Pareto improvements on the table during the bargaining period when they choose the machine's policy.

This phenomenon warrants a few interpretations:

Bet settling As discussed in Section 3.5, a machine implementing a Pareto optimal policy can be viewed as a kind of bet-settling device. If Alice is 90% sure the Four Horsemen will appear tomorrow and Bob is 80% sure they won't, it makes sense for Alice to ask—while bargaining with Bob for the machine's policy—that the machine prioritize her values more if the Four Horsemen arrive tomorrow, in exchange for prioritizing Bob's values more if they don't. Both parties will be happy with this agreement in expectation. As long as it remains possible to redistribute the machine's priorities in a way that resembles an agreeable bet between Alice and Bob, its policy is not yet Pareto optimal. Thus, Theorem 8 can be seen as saying that a Pareto optimality policy goes about settling a bet on the machine's input at each time step, in such a way that no additional bets settlable by the policy are desirable to both players.

Moral realism with Bayesian updating Alternatively, we could take more seriously the interpretation of the weights w^j in Theorem 8 as prior "beliefs" about the value of the made-up latent variable B from Lemma 5 that simultaneously governs (1) how the environment works, and (2) what utility function is "correct" to pursue. This interpretation is a bit unnatural because, even if the original environmental variables S_i were very grounded in physical reality, the abstract variable B in Lemma 5 is merely a fiction conjured up to imply a correlation between "is" and "ought": namely, that either the world *is* governed by $\mathbb{P}^1$, and U^1 *ought* to be optimized, or the world *is* governed by $\mathbb{P}^2$, and U^2 *ought* to be optimized. This occurs even if each of the two players treats their beliefs and utilities completely separately (i.e., even if they apply Bayesian updating only to their beliefs, and keep their utility functions fixed).

Mixing "is" and "ought" in this way is often considered a type error. Nonetheless, many humans report an intuitive sense that there are objective, right-and-wrong answers to moral questions that can be answered by observing the world. If a human is implicitly and approximately acting in a Pareto optimal fashion for a mixture of belief/utility outlooks $D^1, \ldots, D^k$, then the process of "updating" to favor a certain utility function might feel, from the inside, like "finding an answer" to a moral question.

5 Current limitations and future directions

The eventual aim of this work is to facilitate the cooperative development and deployment of advanced AI systems, by simplifying the process of bargaining for shared control of such systems, and by making collaborative outcomes generally easier to implement and more attractive. For this purpose, while Pareto optimality is a desirable condition to aim for, it is not an adequate solution concept on its own. Indeed, the policy "maximize player 1's utility function, without regard for player 2" is Pareto optimal, yet is clearly not the sort of solution one would expect two nations to agree upon. This and at least several other issues must be addressed to build a satisfactory negotiation framework, exhibited below in order of increasing difficulty as estimated by the author.

1. BATNA dominance. In any bargaining situation, each player has a "best alternative to negotiated agreement", or BATNA, that she expects to obtain if the other player chooses not to cooperate. The characterization of Pareto optimality given in Theorem 8 *does not* account for the players' BATNAs. Given a facet of the Pareto boundary, specified by the maximization of a linear function with weights (w^1, w^2), a policy π satisfying Theorem 8 will yield an expectation pair $\left(\mathbb{E}^1[U^1;\pi], \mathbb{E}^2[U^2;\pi]\right)$ lying on that facet. Thus, the bargaining problem has been reduced to choosing an appropriate facet of the Pareto boundary. But suppose not all points on the chosen facet lie above both players' BATNAs. Then, in order to satisfy the individual rationality of the players, the policy *should* target a more specific subset of that facet.

2. Targeting specific expectation pairs. If a specific target value for the expectation pair $\left(\mathbb{E}^1[U^1;\pi], \mathbb{E}^2[U^2;\pi]\right)$ is desired, unless that pair is a vertex of the Pareto region (e.g., perhaps the boundary is curved), the best that following the recursion of Theorem 8 ensures is a point on the same facet. The ability to target a specific pair would solve not only BATNA dominance (1), but also help achieve other fairness or robustness criteria that might arise from bargaining. One approach would be to make a small modification to the players' utility functions to ensure that the resulting Pareto boundary is curved, thereby avoiding this problem at the cost of a tiny utility adjustment. Choosing a simple form for the adjustment that is amenable to formal proof would be a natural next step in this direction.

3. Information trade. Our algorithm implicitly favors whichever player best predicts the machine's input history, given its action history. This makes sense when the players have common knowledge of each other's priors and observations, at which point they have already had the opportunity to update on each other's views and chosen not to. This is unrealistic if Alice knows that Bob has made observations in the past that Alice did not. In that case, Alice will view Bob's beliefs as containing valuable information that ought to shift her prior. She may wish to bargain with Bob for access to that information in order to improve her own ability to optimize the machine's policy. Perhaps she would concede some control over the machine (by reducing her weight, w^1) in exchange for information provided by Bob to improve her beliefs. An efficient procedure to naturally facilitate this sort of exchange would be complimentary Theorem 8. One approach would be to have each player express their posterior as a function of the other's, and use a fixed point theorem to choose a stable pair of posteriors. However, many questions arise about this method when there are multiple fixed points.

4. Learning priors and utility functions. It is notoriously difficult to explicitly specify one's utility function U to a machine, so in practice, one must choose a method enabling the machine to learn the utility function. Cooperative inverse reinforcement learning (CIRL) (Hadfield-Menell et al., 2016) exhibits such a framework, and reduces the problem to solving a POMDP. In CIRL, a human and a robot play a cooperative game wherein both players aim to maximize the human's utility function U, but the robot is uncertain about U and must infer it from the human. Moreover, the human and robot have common knowledge of this situation, so the human may engage in "teaching" behavior to help the robot along. Such dynamics must be accounted for in a satisfactory treatment of negotiation for a machine's priorities. In addition, the players' priors should probably also be learned by a machine in some way rather than explicitly specified.

5. Incentive compatibility. Assuming any particular method for learning players' priors and utility functions, a question arrises as to whether it incentivizes players to represent their beliefs and utilities honestly. For example, Alice may have some incentive to exaggerate her estimation of her BATNA in the positive direction, to motivate Bob to "sweeten the deal" by conceding her a higher priority w^1 in the recursion of Theorem 8. As well, players might also have incentives to alter their reported beliefs in order to exaggerate the degree to which the machine's decisions will affect their utilities. A satisfactory learning method should rule out or otherwise cope with this phenomenon. A great deal of literature already exists on incentive compatibility, as begun by Hurwicz (1972), Myerson (1979), and Myerson and Satterthwaite (1983), which should offer a good start.

6. Naturalized decision theory. The POMDP setting used here is "Cartesian" in that it assumes a clear divide between the machine's inner workings and its environment. This is highly inappropriate when the machine may be copied or simulated; it may wind up in Newcomb-like problems as in Soares and Fallenstein (2015), and very strange cooperative equilibria may exist between its copies, such as in Critch (2016). Instead, one should assume a "naturalized" model of the problem where the machine is part of its environment, as in Fallenstein, Soares, and Taylor (2015). Some attempts have been made to characterize optimal decision-making in a naturalized setting, e.g., by Orseau and Ring (2012), but very few theorems to aid in sequential implementation exist (e.g., no analogue of Proposition 7 is known), except possibly for some self-reflective properties exhibited by Garrabrant's logical inductors (Garrabrant et al., 2016) that might be expanded to exhibit relevance to sequential decision-making. Without a satisfactory model of naturalized decision-making for the machine to follow, the negotiating parties might unwittingly assign the machine a policy vulnerable to Newcomb-like extortions. On the other hand, a satisfactory resolution would not only help to model the machine's situation, but also that of the players themselves during the negotiation phase.

6 Conclusion

Insofar as Theorem 8 is not particularly mathematically sophisticated—it employs only basic facts about convexity and linear algebra—this suggests there may be more low-hanging fruit to be found in the domain of "machine implementable social choice theory". To recapitulate, Theorem 8 represents two deviations from the intuition of naïve utility aggregation: to achieve Pareto optimality for players with differing beliefs, a machine must (1) use each player's own beliefs in evaluating how well an action will serve that player's utility function, and (2) shift the relative priority it assigns to each player's expected utilities over time, by a factor proportional to how well that player's beliefs predict the machine's inputs.

As a final remark, consider that social choice theory and bargaining theory were both pioneered during the Cold War, when it was particularly compelling to understand the potential for cooperation between human institutions that might behave competitively. In the coming decades, machine intelligences will likely bring many new challenges for cooperation, as well as new means to cooperate, and new reasons to do so. As such, new technical aspects of social choice and bargaining, along the lines of this paper, will likely continue to emerge. In particular, the problems outlined in Section 5 represent areas particularly promising for facilitating cooperative outcomes in the deployment of advanced AI systems, and the present author is seeking collaborations to address them.

References

Abbeel, Pieter and Andrew Y Ng (2004). "Apprenticeship learning via inverse reinforcement learning". In: *Proceedings of the twenty-first international conference on Machine learning*. ACM, p. 1.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman (2016). "Racing to the precipice: a model of artificial intelligence development". In: *AI & SOCIETY* 31.2, pp. 201–206.

Baum, Seth D (2016). "On the promotion of safe and socially beneficial artificial intelligence". In: *AI & SOCIETY*, pp. 1–9.

Bostrom, Nick (2014). *Superintelligence: Paths, dangers, strategies*. OUP Oxford.

Critch, Andrew (2016). "Parametric Bounded Lob's Theorem and Robust Cooperation of Bounded Agents". In: *arXiv preprint arXiv:1602.04184*.

Darwiche, Adnan (2009). *Modeling and reasoning with Bayesian networks (Chapter 4)*. Cambridge University Press.

Fallenstein, Benja, Nate Soares, and Jessica Taylor (2015). "Reflective variants of Solomonoff induction and AIXI". In: *International Conference on Artificial General Intelligence*. Springer, pp. 60–69.

Gbor, Zoltn, Zsolt Kalmr, and Csaba Szepesvri (1998). "Multi-criteria Reinforcement Learning." In: *ICML*. Vol. 98, pp. 197–205.

Garrabrant, Scott, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor (2016). "Logical Induction". In: *arXiv preprint arXiv:1609.03543*.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell (2016). *Cooperative Inverse Reinforcement Learning*.

Harsanyi, John C (1980). "Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility". In: *Essays on Ethics, Social Behavior, and Scientific Explanation*. Springer, pp. 6–23.

Hurwicz, Leonid (1972). "On informationally decentralized systems". In: *Decision and organization*.

Hutter, Marcus (2003). *A Gentle Introduction to The Universal Algorithmic Agent {AIXI}*.

Myerson, Roger B (1979). “Incentive compatibility and the bargaining problem”. In: *Econometrica: journal of the Econometric Society*, pp. 61–73.

— (2013). *Game theory*. Harvard university press.

Myerson, Roger B and Mark A Satterthwaite (1983). “Efficient mechanisms for bilateral trading”. In: *Journal of economic theory* 29.2, pp. 265–281.

Nash, John F (1950). “The bargaining problem”. In: *Econometrica: Journal of the Econometric Society*, pp. 155–162.

Ng, Andrew Y, Stuart J Russell, et al. (2000). “Algorithms for inverse reinforcement learning.” In: *Icml*, pp. 663–670.

Orseau, Laurent and Mark Ring (2012). “Space-Time embedded intelligence”. In: *International Conference on Artificial General Intelligence*. Springer, pp. 209–218.

Pearl, Judea (2009). *Causality*. Cambridge university press.

Roijers, Diederik M, Shimon Whiteson, and Frans A Oliehoek (2015). “Point-based planning for multi-objective POMDPs”. In: *IJCAI 2015: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1666–1672.

Russell, Stuart (1998). “Learning agents for uncertain environments”. In: *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, pp. 101–103.

Russell, Stuart, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards (2003). *Artificial intelligence: a modern approach (Chapter 17.1)*. Vol. 2. Prentice hall Upper Saddle River.

Shoham, Yoav and Kevin Leyton-Brown (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.

Soares, Nate and Benja Fallenstein (2015). “Toward idealized decision theory”. In: *arXiv preprint arXiv:1507.01986*.

Soh, Harold and Yiannis Demiris (2011). “Evolving policies for multi-reward partially observable Markov decision processes (MR-POMDPs)”. In: *Proceedings of the 13th annual conference on Genetic and evolutionary computation*. ACM, pp. 713–720.

Tzeng, Gwo-Hshiung and Jih-Jeng Huang (2011). *Multiple attribute decision making: methods and applications*. CRC press.

Wang, Weijia (2014). “Multi-objective sequential decision making”. PhD thesis. Universit Paris Sud-Paris XI.

Wray, Kyle Hollins and Shlomo Zilberstein (2015). “Multi-objective POMDPs with lexicographic reward preferences”. In: *Proceedings of the 24th International Joint Conference of Artificial Intelligence (IJCAI)*, pp. 1719–1725.

Zhang, Chongjie and Julie A Shah (2014). “Fairness in multi-agent sequential decision-making”. In: *Advances in Neural Information Processing Systems*, pp. 2636–2644.